SOME COMPARATIVE STUDIES ON THE CONSTRUCTION OF COMPOSITE REGIONAL INDICES

R. N. Chattopadhyay and M. N. Pal*

1. The Objective

1.1 The purpose of this paper is to show that the composite indices as are constructed under the usual criterion of mathematical optimality need not necessarily be the one that is most suited for a proper identification of a particular formal regional configuration. The logic of construction of a mathematically optimal composite index can be summarised as follows :

Let a composite characteristic be reflected in n variables, each varying over N spatial units of observation. Let these variables be represented by the vector

$$X_k = (x_{1k}, x_{2k}, ..., x_{nk}) ; k = 1, 2, ..., N,$$

where X_k can be considered as a point (kth point) in the Euclidean n-space. As a common composite characteristic is influenenced by all n variables x_i, $i = 1, 2, ... n$, one can expect a high degree of correlation r_{ij} between variables x_i and x_j ; $i, j = 1, 2, ... n$. If the correlation r_{ij}'s were all perfect (*i.e.* $| r_{ij} | = 1$), the points would lie on a curve which would approximate to a straight line when the statistical distributions of variables are similar (similar distributions are essential for a validity of hypothesis on the linear type of relationship between variables ; these linear correlation coefficients are computed in various correlation matrices). Then the order of the points of projection on this line would enable us to rank the spatial units in order of the composite characteristic under determination. In practice, however, the correlations are not perfect. We, therefore, determine, in accordance with the usual procedure in analogous cases, the straight line of closest fit to the cluster of N points and rank the spatial units by a related equation called ,'composite index formula", giving the order of points of projection onto this line of closest fit. Kendall [1] constructed a composite agricultural productivity index for his studies by counties in

* R. N. Chattopadhyay is associated with the Indian Institute of Technology, Kharagpur, while M. N. Pal is associated with the Indian Statistical Institute, New Delhi.

England as early as in 1939. We shall refer to this composite index formula as "Kendall's formula" in our subsequent discussions. This index has the highest aggregate correlation β, with its constituent variables x_1 to x_n. By aggregate correlation we mean that the coefficient β is given by

$$\beta = \sqrt{\left(\frac{1}{n}\sum_{i=1}^{n} r_i^2\right)},$$

where r_i is the usual product moment correlation coefficient between an index and its constituent variable $x_i, i = 1, 2, \cdots n$. Kendall's index formula is optimal in the sense that the representativeness of variables in the index, measured by the aggregate correlation coefficient β, is highest.

1.2. We shall show that even though β is maximum in Kendall's formula, maximising the aggregate representativeness, the specific representativeness of certain constituent variables (r_i's) may be so low that it can not be suitable for the identification of a composite spatial configuration involving all n variables. In such a situation, one may have to prescribe certain minimum standard for each specific representativeness. If this minimum standard is not satisfied in the Kendall's formula, the corresponding index has to be rejected. Alternatively Pal [3, 4] derived a formula as early as in 1963 for two constituent variables and generalized it subsequently to the case of n constituent variables; this formula is based on the principale of equal specific representativeness by every constituent variables. That is, we have $r_1 = r_2 = \ldots = r_n = r$, say, and hence the aggregate correlation $\beta = r$ obtained a high value of aggregate correlation β, one may have to reject Kendall's formula on grounds of specific representativeness problem, whereas Pal's formula does not suffer from such a problem. As Kendall's formula is optimal, the value of β in Pal's formula cannot exceed that in Kendall's. As such there is a certain loss in terms of aggregate representativeness in Pal's formula. If this loss is not statistically significant, Pal's formula serves better than the optimal Kendall's formula in the study of composite spatial configuration. Even otherwise, this loss may not weigh heavily in many situations, because of the gain achieved in specific representativeness of all constituent variables in Pal's index. We shall illustrate these points by examples taken from the Ph.D. dissertation on "Regionalisation and Regional Development of the Coal-Steel Belt of India". by R. N. Chattopadhyay.

1.3. Incidentally, we shall examine another important aspect of

statistical applications that did not receive adequate attention in spatial analysis. The statistical regression techniques as applied in spatial analysis over a certain universe of study stress on finding the global relations $r_{i,j}$'s, taking count of all spatial observations N, between two (or more) spatial variables, x_i, x_j etc. This so-called global relations may be at times misleading or considerably dampened if the choice of universe is such that there is a possibility of mixing *conformal* (conforming to the pattern depicted by an overwhelming majority of N spatial observations) and *non-conformal* (not conforming to the majority pattern and at times just a reverse pattern may be reflected by a minority group of residual spatial observatation) type of relations in the totality of all spatial observations. This problem of mixing up conformal and non-conformal relations is more pronounced in a large heterogeneous universe of study, wherein a true relation is dampened considerably because of its heterogeneity. We shall illustrate a practical procedure of depicting conformal relations and show how the analysis based on conformal relations could be extended to the residual non-conformal cases of spatial observations.

2. Empirical Background :

2.1. The starting point of this paper is the consideration of values under three initial indices, x_1, x_2 and x_3 as the raw data for our analysis over an universe of study containing fortynine (49) districts in and around the Coal-Steel Belt of India. The variables x_1, x_2 and x_3 are referring to (i) agricultural activities,, (ii) mining activities, and (iii) non-agromining activities in this analysis. The constructions of these variables is not the matter of interest of this paper. However, we would describe them very briefly in the following lines to get an impression on the nature of data to be used at the starting point.

2.2. The agricultural index x_1 was derived by Kendall's formula applied on following four initial variables expressed in some suitable mathematical forms : (1.1) agricultural labour productivity, (1.2) agricultural land productivity, (1.3) extent of irrigation and (1.4) agricultural wage rate. The mining variable x_2 is the values of aggregate mining outputs in 1966 expressed in some suitable mathematical form. The non-agromining index x_3 was derived by Kendall's formula applied on four sub-indices related to (3.1) secondary activities, (3.2) tertiary activities, (3.3) urbanisation, and (3.4) literacy rates, while sub-indices (3.1), (3.2) and (3.3) had been derived similarly by Kendall's formula

on the basis of following variables expressed in some suitable mathematical forms: (3.11) share of income in manufacturing, (3.12) areal concentration of secondary labour, (3.13) sectoral concentration of secondary labour, (3.14) areal concentration of large factories, (3.15) areal concentration of all factories, (3.16) areal concentration of factory workers and (3.17) share of large factory employment in secondary labour; (3.21) share of income in tertiary activities, (3.22) areal concentration of tertiary labour, (3.23) sectoral concentration of tertiary labour, and (3.24) areal concentration of transport arteries; (3.31) share of urban population, (3.32) areal concentration of urban population, (3.33) average size of a town and (3.34) areal concentration of city-population (each city has 50,000 or more inhabitants).

Unless otherwise stated, all initial variables are based on data taken from the following sources: (i) 1961 Census, (ii) Annual survey of Industries, (iii) Labour Bureau, (iv) National Council of Applied Economic Research, (v) Central Statistical Organization, (vi) Indian Bureau of Mines, (vii) Directorates of Metals and mines for five eastern States of India, (viii) Chief Inspectorate of factories for five eastern States of India. The aggregate representativeness as reflected in the different indices and sub-indices by the constituent variables were high enough to be acceptable to us; thus these values of β's are as given below with the specific representativeness of constituent variables indicated in the vector following the data on β's.

$\beta(x_1) = 0.828$; (0.871, 0.880, 0.689, 0.857)
$\beta(x_2) = 1.000$; (1.000)
$\beta(x_{3\cdot1}) = 0.850$; (0.584, 0.822, 0.878, 0.950, 0.910, 0.971, 0.769)
$\beta(x_{3\cdot2}) = 0.930$; (0.922, 0.937, 0.971, 0.890)
$\beta(x_{3\cdot3}) = 0.933$; (0.950, 0.941, 0.920, 0.920)
$\beta(x_{3\cdot4}) = 1.000$; (1.000)
$\beta(x_3) = 0.942$; (0.944, 0.939, 0.963, 0.916).

All specific representativeness were high enough to be above an acceptable minimum standard of over 0.500 and as such Kendall's formula is not questioned for acceptance of x_1, x_2 and x_3 as initial variables for our analysis.

3. Logical Development of the Methods of Analysis:

3.1. We are now faced with the problem of finding a composite activity index based on variables x_1: agricultural activity index, x_2:

mining activity index, and x_3 : index of non-agromining activities. The total number of spatial units of observations is 49 in our universe of study. If we compute a correlation matrix over these 49 observations, obtaining the correlation coefficients r_{ij}'s between variables x_i and x_j; $i, j = 1, 2, 3$, we get negative r_{ij}'s everywhere except for the diagonal elements r_{ii}, $i = 1, 2, 3$.

TABLE (3.11) : *Correlation Matrix for N=49 : r_{ij} (49)*

	x_1	x_2	x_3
x_1	1.0000	−0.0233	−0.1093
x_2	−0.0233	1.0000	−0.2668
x_3	−0.1093	−0.2668	1.0000

This correlation matrix is presented in table (3.11) below. These negative values on r_{ij} (49)'s seem to be paradoxically opposite to our usual expectations of having positive inter-relation between agricultural, mining and non-agromining variables. If one examines now the raw data one discovers that a metropolitan district like Calcutta has the minimum zero values for x_1 and x_2 and has, at the same time, the maximum value for x_3; this local peculiarity is understandable. Omitting this peculiar observation from our computation of correlation matrix, we get the following values of correlation coefficients as given in table (3.12).

TABLE (3.12) : *Correlation Matrix for N=48 : r_{ij}(48)*

	x_1	x_2	x_3
x_1	1.0000	−0.1583	0.4637
x_2	−0.1583	1.0000	−0.1705
x_3	0.4637	−0.1705	1.0000

It should be kept in mind that correlation matrices are symmetric in nature so that we have always $r_{ij} = r_{ji}$. Now r_{21} (48) and r_{23} (48) still retain the negative values while r_{13}(48) changes to a considerably high positive value of 0.4637. This means that the mining variable has still not depicted a logically conformal pattern in relation to other two variables, while the omission of Calcutta gives a conformal relation between agricultural and non-agromining variables. On examination again of the raw data, we find that nine districts do not have any

mining activity at all with values as equal to zero for x_3, while their positions in respect to x_1 and x_2 are on the higher side on the respective scales of values. Omitting ten non-mining districts (including Calcutta) we tried to evaluate the inter-relations of variables for $N=39$. The results of this recomputations are shown in table (3.13).

TABLE (1.13): *Correlation Matrix for* $N=39: r_{ij}(39)$

	x_1	x_2	x_3
x_1	1.0000	0.0466	0.5059
x_2	0.0466	1.0000	0.0745
x_3	0.5059	0.0745	1.0000

Now all $r_{ij}(39)$'s turn out to be positive, yet correlation $r_{21}(39)$ and $r_{23}(39)$ are not very different from the stage of no correlations. We searched for other non-conformal observations and found that the pattern depicted by two districts gave reverse pattern to that brought out through remaining 37 observation. As such we decided to omit these two districts also for our computation of the correlation matrix. Finally we get the correlation matrix as presented here in table (3.14).

TABLE (3.14): *Correlation Matrix for* $N=37: r_{ij}(37)$

	x_1	x_2	x_3
x_1	1.0000	0.1865	0.5290
x_2	0.1865	1.0000	0.1915
x_3	0.5290	0.1915	1.0000

It could be noticed that $r_{ij}(37)$'s are considerably higher than corresponding values of $r_{ij}(39)$, $r_{ij}(48)$ and $r_{ij}(49)$. Also it could be noticed that the mining variable has considerably improved its relations with each of other two variables. As such the construction of composite activity index was considered on the basis of the correlation matrix, valid for 37 conformal observations.

3.2. *Composite Activity Index by Kendall's formula*: Starting from the correlation matrix of table (3.14), the "composite activity index" was computed by Kendall's formula and given below in equation (1)

$$Y = 0.54311\, x_1 + 0.06578\, x_2 + 0.42983\, x_3 \qquad (1)$$

Here the mean of Y over 37 observations is unity. Its aggregate representativeness β is given below, followed by a vector of the specific representativeness of constituent variables.

$$\beta(Y) = 0.7395\ ;\ (0.8348,\ 0.4933,\ 0.8369) \qquad \ldots \qquad (2)$$

Now the equation (1) can safely be used to evaluate Y_i's for first 37 districts and also for two other districts ; none of these 39 districts have zero values for x_1, x_2 and x_3. Trouble comes for subsequent nine observations, wherein there are zero values for x_2 and also for the district of Calcutta for which both x_1 and x_2 show up zero values. A linear equation of the type given in equation (1) will pull down the ranks because of the consequential zero-effects of the coefficients of x_2 over districts with serial numbers 40 to 48 and that of the coefficients of x_1 and x_2 for the district of Calcutta. So for the calculations of Y values for these ten districts we determined the following regression equations (3) and (4) by the usual least square methods.

$$Y' = 0.03038 + 0.64140\, x_1 + 0.37808\, x_3 \qquad \ldots \qquad (3)$$

This equation was determined from first 39 observations giving a multiple correlation coefficient $R = 0.844$; the equation (3) has been used for the calculation of Y values for the districts with serial numbers 40 to 48 ; these values are recorded under column (5) of the table (1) of Appendix A. For the calculation of Y value for the district of Calcutta (serial number 49) we used the following regression equation (4), determined from preceding 48 observations, giving a value of correlation coefficient as equal to 0.975 ; this Y value for Calcutta is :

$$Y'' = 0.42115 + 0.65488\, x_3 \qquad \ldots \qquad (4)$$

The greater magnitude of correlation coefficients for equations (3) and (4) than the β-value for equation (1) indicate the better predictive power of equations (3) and (4) for Y values of non-conformal districts (with serial numbes 40 to 49). Thus, finally we have all computed values of composite activity index Y by Kendall's formula and extended regression equations.

3.3. *Composite Activity Index by Pal's formula* : It could be noted from the vector given after equation (2) of para 3.2, that the value of 0.49 showing the specific representativeness of x_2 (mining variable) in composite activity index Y has not attained the minimum acceptable value of even 0.500, when the value of 0.74 for aggregate representativeness is considerably above this standard. This means that even

though our problem is to depict a formal configuration of composite activities in which mining activity must get a duly appropriate representation, depending upon the requirements for the identification of mining (coal steel) belt, the mining variable did not get sufficient representation in the Kendall's composite activity index. As such, we propose to compute an alternative composite activity index in accordance with the methods of Pal's formula. Again starting from the correlation matrix of table (3.14) the "composite activity index" was computed by Pal's formula and given below in equation (5).

$$Z = 0.49051\, x_1 + 0.14161\, x_2 + 0.38145\, x_3 \qquad \ldots \quad (5)$$

Here the mean $\bar{Z} = 1$ as calculated over the 37 observations used in the correlation matrix. Its aggregative representativeness β is given below which is also equal to each specific representativeness in this method.

$$\beta(Z) = 0.725 \qquad \ldots \quad (6)$$

Following the same scheme as used in para 3.2, the equation (5) was used to evaluate Z's for first 39 districts; in these 39 districts none of x_1, x_2 and x_3 have zero-values. For calculations of Z-values for nine other districts the regression equation (7) was determined by usual least square method.

$$Z' = 0.14620 + 0.53673\, x_1 + 0.36571\, x_3 \qquad \ldots \quad (7)$$

This equation (7) was determined from first 39 observations, giving a multiple correlation coefficient $R = 0.776$; this equation (7) has been used for the calculations of Z-values for the nine districts. For the calculation of Z-value for the Z-last district of Calcutta we used the following regression equation (8), determined from preceding 48 observations, having a value of correlation coefficient as equal to 0.775; this Z-value for Calcutta is:

$$Z'' = 0.46136 + 0.60876\, x_3 \qquad \ldots \quad (8)$$

Here also the higher magnitude of correlation coefficients for equations (7) and (8) than the β-value for equation (5) indicate the better predictive powers of equations (7) and (8) for Z-values of non-conformal districts. Thus, here also, we have all computed values of composite activity index Z and extended regression equations.

3.4. *Comparisons of Specific Representativeness of Variables in the two Alternatives*: The specific representativeness of variables x_1, x_2 and x_3 under the two alternatives are available in equations (2) and (6). For the alternative of Pal's index the aggregate representativeness $\beta(Z)=0.725$

given in equation (6) stands for each specific representativeness. The following table (3.41) gives the results of Fisher's Z-test on correlation coefficient. This table shows, with statistical rigour, that the specific representativeness of x_8 in Kendall's index is significantly below : the

TABLE (3.41) : *Comparisons of Specific Representativeness in two Alternative Indices by Fisher's Z-test.*

	r_i = specific representativeness in Kendall's formula	$Z_i = \frac{1}{2}\log_e\frac{(1+r_i)}{(1-r_i)}$	$Z' = \frac{1}{2}\log_e\frac{1+\beta(Z)}{1-\beta(Z)}$, $\beta(Z)=0.725 = r_1=r_2=r_3$ in Pal's formula	test function $t = \lvert(Z_i - Z')\rvert \sqrt{(N-3)}$, $N=37$	Remarks
col. 1	col. 2	col. 3	col. 4	col. 5	col. 6
1.	0.8348	1.20378	0.91811	1.6657	not significant
2.	0.4933	0.54041	0.91811	2.2024	significant at 5 p.c. level of chance
3.	0.8569	1.21074	0.91811	1.7063	not significant

N.B. : The test function t is normally distributed.

specific representativeness of x_8 in Pal's index, whereas the specific representativeness of x_1 or x_3 is not different significantly in the two alternative indices Y and Z. Again the relation between Y and Z is as high as to show the value of correlation coefficient to be equal to 0.957 for $N=37$ and equal to 0.975 for $N=49$. The regression equation of Y and Z values determined from all 49 observations by usual least square method, is given below in equation (9).

$$Y = 0.02297 + 0.98390\, Z\,;$$

$$N = 49, \qquad r = 0.975. \qquad \ldots \quad (9)$$

The near-unity value of correlation coefficient r for the regression equation (9) shows that the indices Y and Z are equally acceptable from the statistical viewpoint, while the results of table (3.41) shows that Y is not acceptable on grounds of under-representation of mining variable x_8 in Y. Z does not suffer from such limitations and is considered most useful between the two.

3.5. *Concluding Remarks on the Results*: As Y and Z have close agreement with high correlation coefficient $r = 0.975$ between them, we get similar ranking of districts by both Y and Z in many areas. Major departure occurs in respect of the Z-values in mining districts, as expected. It could be observed that the mining belt comprising of ten districts could be identified only in Z. A rise to generally high values in Z as compared to corresponding Y-values can be noticed in these districts. Thus Y-values under-represents the intensity of mining activity in these districts. The core of this mining areas is at Dhanbad and Burdwan districts. These two districts show up with very high values in Z definitely—with a considerable rise over corresponding Y-values. Other areas show expectedly almost similar rankings in both Y and Z-values. Poor intensities of activities are observed on either side of this mining districts as identified above. Presence of mining activity may be there in these adjoining areas, but this activity has not yet geared to other allied activities to generate a significant boost. For example, the district of Hazaribagh has substantial mining activity, but it is not geared to other non-agromining activities so that its rank remain medium. Other important area that has been identified in this study is Calcutta and its surronnding districts. The presence of very high values in both Y and Z could be accounted for by the presence of all kinds of industrial activities in these districts including Burdwan. So Burdwan is not merely a mining district. On either side of this industrially advanced area there lies two areas of agricultural importance, one in the north-east with Nadia, Murshidabad and Birbhum districts having high values in both Y and Z, and the other in the south-west with Midnapur and Bankura districts with medium Z-values but high Y-values. Besides, we have a coastal agricultural belt extending from the district of Visakhapatnam to the district of Balasore. Here agricultural activities are of mixed kind. Thus high values at Ganjam district can be accounted for by the existence of a rich rice bowl, while the high value of Cuttack district is largely due to the urbanisation in this area and to some extent due to the presence of some agricultural and mining activities. Sahabad district in the far north show up with high values for the presence of factories based on mining activities (ACC cement factory is located in this district). The adjoining Patna district is also showing high values mainly because of the presence of an urban centre by this name. However, inspite of the general agreement in rankings by Y and Z values, the identification of mining belt by only Z-values is the special gain in this analysis by composite indices.

References :

1. Kendall, M. G. "The Geographical Distribution of Crop Productivity in England." *Journal of the Royal Statistical Society*, Vol. 102, 1939, pp. 21-48.

2. Pal, M. N. "Information System and Regional Development in India." *United Nations Research Institute for Social Development*. Geneva, UNRISD/70/C. 74, GE 71-3383, October 1970.

3. Pal, M. N. "Zur Berechnung eines Kombinierten Konzentrationsindexes" *Raumforschung und Raumordung*, Vol. 21, 1963, pp. 87-93.

4. Pal, M. N. "Quantitative Techniques for Regional Planning," Paper for the Conference of the *Commission on Quantitative Methods, International Geographical Union*, held at the London School of Economics, London, 18-20 August, 1969. *Published in Indian Journal of Regional Science*, Vol. 3, No. 1, 1971, pp. 1-33.